Basic course on regression modelling

Mervi Eerola

Turku Center of Statistics

24-25.9 and 1.10.2015

Outline of the course

1. Basic principles of statistical inference
2. Linear regression: the basics
3. Linear regression: modifying the model
4. Logistic regression

Gelman, A. & Hill, J. (2007): Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research. Cambridge University Press.

Other transformations

- On the log scale, a difference in earnings between \$5000 and \$10000 is equivalent to a difference between \$40000 and \$80000 which may not be realistic
- Sometimes a square root transformation is used instead but the coefficients do not have a natural interpretation in terms of either absolute or relative change, so for models with interpretative purpose it should not used
- In the earnings -example, those who had 0 earnings had to be excluded from the log model. A separate model for those with no earnings and those having earnings could have been postulated
- It is important to check the range of variation with scatterplots and histograms before modelling

Discrete predictors and dummy variates

- In general, it is not wise to discretize a continuous predictor, and especially not to a binary covariate
- Sometimes the interpretation of a continuous predictor is 'multidimensional' and discretization helps interpretation
- Example (Child's score): Maternal employment in the first three years of child's life (4-point ordered scale)
  - `mother.work=1`: did not work at all
  - `mother.work=2`: worked in second or third year
  - `mother.work=3`: worked part-time
  - `mother.work=4`: worked full-time
- Note: this model allows *different averages* in different working categories

Discrete predictors

- One category must be chosen as a *reference* category, other coefficients are interpreted as *deviations* from it
- Example (Child's score). The reference category is the first ('mother did not work at all')
- For example, the predicted test score for children whose mother did not work is 82 and for children whose mother worked part-time is $82 + 11.5$

```
 > display(lm(formula = kid.score ~as.factor(mom.work), data=kid=ic
                                 coef.est coef.se
(Intercept)                82.0      2.3
as.factor(mom.work)2        3.8      3.1
as.factor(mom.work)3       11.5      3.6
as.factor(mom.work)4        5.2      2.7
---
n = 434, k = 4
residual sd = 20.2, R-Squared = 0.02
```

Identifiability

- If the model has parameters that cannot be estimated uniquely, it is said to be *nonidentifiable*
- Including all categories of `mother.work` would cause nonidentifiability
- In general, if a categorical predictor has J categories with separate indicators, only $J - 1$ can be estimated

Identifiability

- Nonidentifiability arises also if predictors are highly or even perfectly correlated (collinearity)
- Standard errors of the coefficients are then very large and the model is useless for predictive purposes
- It is usually wise to start from simple models and add predictors to understand their marginal and joint influence

Guidelines for model building

- Include all covariates that are *a priori* considered to have important predictive value for the outcome
- Some of them may be included as sums or averages (total scores)
- Include interactions for predictors that have a large main effect
- Decision to exclude/include a predictor: evaluate its expected sign and statistical significance, especially if the sign is not as expected but significant (why?)
- Predictors that are not significant but have an expected sign, and have a theoretical meaning, can be included

4. Logistic regression

1. Interpreting logistic regression
2. Logistic regression with interaction
3. Diagnostics
4. Average predictive comparisons

Logistic regression

- Logistic regression is a model for binary outcomes (yes/no, alive/dead, on/off etc.)
- Example (Political preference given income):
  $y = 1$ (voted Bush in 1992), $y = 0$ (voted Clinton)

```
fitt=glm(vote~income,family=binomial(link="logit"))
display(fitt)
            coef.est coef.se
(Intercept) 0.02     0.03
     income 0.33     0.01
---
  n = 33140, k = 2
  residual deviance = 38367.4, null deviance = 39291.1 (difference
```

- The model predicts that those with higher income vote Bush

The logistic model

- The linear regression model $y = X_i\beta + \epsilon_i$ cannot be used to model outcomes of 0 and 1
- Instead, we model the probability that $y = 1$ using the logit function to transform the range (0,1) of probability values to the values in $(-\infty, +\infty)$

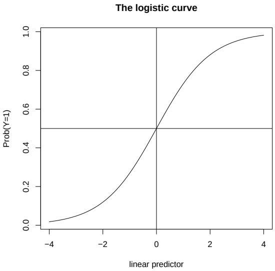
$$\text{For } P(y_i = 1) = p_i,\ \text{logit}(p_i) = X_i\beta,$$

  where the logit transformation is $logit(p) = log(p/(1-p))$
- Another way to write the model is

$$P(y_i = 1) = logit^{-1}(X_i\beta) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$$

  where the inverse logit function maps the continuous values of the linear predictor $X_i\beta$ to the range (0,1) of probabilities

The logistic curve



Interpretation of the logistic relationship

- The logit curve is not a straight line unlike the regression line; therefore the expected difference in the outcome due to a unit change in the predictor is *not constant*
- To interpret the coefficients requires to choose *where* to evaluate changes in x
- "Divide by 4-rule: The logistic curve is steepest around the point $\alpha + \beta x = 0$ where $\text{logit}^{-1}(\alpha + \beta x) = 0.5$
- The slope is maximized at the derivative of that point and takes the value $\beta exp(0)/(1 + exp(0))^2 = \beta/4$
- Therefore $\beta/4$ is an approximation of the maximum difference in the probability of $y = 1$ due to a unit change in x

Interpretation of the coefficients

- Example (Political preference given income):

$$P(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$$

- "Divide by 4-rule yields $0.33/4 = 0.08$, that is, one unit change in income category corresponds to 8% change in the probability of voting Bush
- Interpretation of the intercept at $x = 0$ is not possible on the 1-5 scale but evaluation at the central value (mean income) $\bar{x} = 3.1$ gives

$$P(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot 3.1) = 0.13.$$

This is not far from the divide-by-four approximation 0.08 because most of the data is near the central value.

Interpretation of the coefficients as odds ratios

- For an outcome with probability p, the *odds* of it is

$$p/(1-p)$$

- The ratio of odds of two outcomes with probabilities p_1 and p_2 is called the *odds ratio* and is of the form

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

- For example, a 2-fold odds ratio corresponds to a change from $p = 0.33$ to $p = 0.5$ on the probability scale because $(0.5/0.5)/(0.33/0.67) = 2.03$ but also from $p = 0.47$ to $p = 0.65$ because $(0.65/0.35)/(0.47/0.53) = 2.09$ so odds ratios are related to particular comparisons of events

Interpretation of the coefficients as odds ratios

- Exponentiated logistic regression coefficients can be interpreted as odds ratios because

$$\frac{P(y_i = 1|x)}{P(y_i = 0|x)}) = e^{\beta x}$$

- If, for example, $\beta = 0.2$ then a unit difference in x corresponds to a 22% change in the odds because $e^{\beta} = 1.22$. This corresponds to, for example, a change from $p = 0.50$ to $p = 0.55$ on the probability scale.

Uncertainty in the coefficients

- As in linear regression models, coefficient estimates within 2sd of $\hat{\beta}$ are consistent with data and those 2sd away from 0 are statistically significant
- The sign or significance of the intercept is not interesting
- An unobserved (future) observation $\tilde{y}$ has a *predictive probability*

$$\tilde{p} = P(\tilde{y}_i = 1) = logit^{-1}1(\tilde{X}_i\beta) = \frac{e^{\tilde{X}_i\beta}}{1 + e^{\tilde{X}_i\beta}}$$

Predictive probability: example

- Example (Political preference given income): A voter not in the survey with income level 5 has the predicted probability of voting Bush

$$P(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot 5) = 0.55$$

- Note: this is the *prediction* of the probability of 'voting Bush', not the probability of the outcome itself

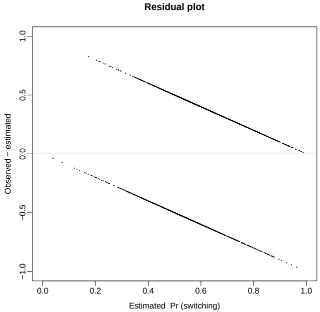
Evaluating and comparing logistic regressions

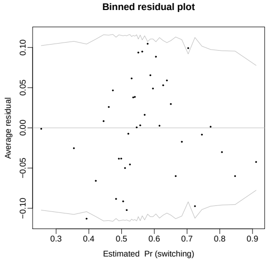
- Residual analysis is in logistic models more complicated than in linear regression models
- The data are discrete (0,1), and so are the residuals, and as such usually not very useful

$$\text{residual}_i = y - \text{logit}^{-1}(X_i\beta)$$

- *Binned residuals*: divide the data into even categories (bins) based on their fitted values and plot the average residual vs average fitted value $X\hat{\beta}$ for each bin
- If most of the binned residuals fall within the $\pm 2sd$ bounds, the model fits reasonably well to the data
- Binning and plotting the residuals against individual predictors can reveal unexpected patterns in data and suggest for transformations, as in linear regression

Logistic residuals and binned residuals





Error rate

- The error rate is the proportion of cases for which the guess: $y_i = 1$ if $\text{logit}^{-1}(X_i\beta) > 0.5$ or $y_i = 0$ if $\text{logit}^{-1}(X_i\beta) < 0.5$ is wrong. It should always be $< 1/2$
- Error rate of the null model: each y_i has the same probability $p = \sum_i y_i/n$ so the error rate is the minimum of $(p, 1-p)$

Error rate: example

- Example (Well-switching): From those with $\text{logit}^{-1}(X_i\beta) > 0.5$, 58% were switchers, 42% not, so the error rate of the null model is $\min(58, 42) = 42\%$
- The final error rate of the models is 36%, indicating that 64% of switchers could be classified correctly with the logistic model, with an improvement of only 6%
- However, most of the data lies near the mean level of the predictors so that the 'null model' (same probability) predicts them well

Deviance

- In discrete data models such as the logistic, the squared error is not a good measure of model fit
- In such models, *deviance* corresponds to the residual variance σ^2 as a summary of error or misfit; the smaller the better fit
- A predictor of pure random noise would decrease deviance by 1 on average
- A meaningful predictor is expected to decrease deviance much more, and k predictors at least by more than k
- Only comparisons *between* models are meaningful: $D = -2log(L_2/L_1)$ where the likelihood L_2 of the larger model 2 is compared to the likelihood of smaller model 1 (at the extreme the null model L_0)

Average predictive comparisons

- As noted earlier, a unit difference in x does not imply a constant change in y in the logistic model so that the regression coefficients cannot be intepreted on the data scale
- Odds and odds ratios are often also difficult to interpret
- Therefore, more useful is to summarize the results as expected or average differences between some interesting groups

Average predictive comparisons: example

- Example (Well-switching): Average predictive difference in P(switch) between households that are next to, or 100m from the nearest safe well (dist100=0 vs dist100=1)
- Write the difference in probabilities δ as a function of `arsenic` and `educ`

$$\begin{aligned}\delta(\text{ars,educ}) = \quad & \text{logit}^{-1}(-1.21 - 0.9 \cdot 1 + 0.47 \cdot \text{ars} + 0.17 \cdot \text{educ}) \\ & -\text{logit}^{-1}(-1.21 - 0.9 \cdot 0 + 0.47 \cdot \text{ars} + 0.17 \cdot \text{educ})\end{aligned}$$

- The average predictive difference is then calculated over the whole sample with the original values in the other predictors $\frac{1}{n}\sum_{i=1}^{n} \delta(arsenic, educ)$